

ACADEMIC TALENT MODEL BASED ON HUMAN RESOURCE DATA MART

Mahani Saron¹, Zulaiha Ali Othman²

**Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM), Malaysia*

¹Email: anie1902@yahoo.com

²Email: zao@ukm.my

Abstract: *In higher education such as university, academic is becoming major asset. The performance of academic has become a yardstick of university performance. Therefore it's important to know the talent of academicians in their university, so that the management can plan for enhancing the academic talent using human resource data. Therefore, this research aims to develop an academic talent model using data mining based on several related human resource systems. In the case study, we used 7 human resource systems in one of Government Universities in Malaysia. This study shows how automated human talent data mart is developed to get the most important attributes of academic talent from 15 different tables like demographic data, publications, supervision, conferences, research, and others. Apart from the talent attribute collected, the forecasting talent academician model developed using the classification technique involving 14 classification algorithm in the experiment for example J48, Random Forest, BayesNet, Multilayer perceptron, JRip and others. Several experiments are conducted to get the highest accuracy by applying discretization process, dividing the data set in the different interval year (1,2,3,4, no interval) and also changing the number of classes from 24 to 6 and 4. The best model is obtained 87.47% accuracy using data set interval 4 years and 4 classes with J48 algorithm.*

Keywords: *Academician database, Classification, Data mart, Talent, Forecasting.*

I. INTRODUCTION

Department of Human Resource Management (HRM) is working in employee-related activities in an organization[1]. Human resources are limited to a particular organization, it is important to be managed effectively in helping the organization towards excellence. HRM is currently having to deal with many challenges such as globalization, to increase the income of the organization, technology changes, manage intellectual capital and a challenge to change [2]. The intellectual capital is one of the challenges faced by the HRM. Finding, developing and retaining talent is the main concern for human resource

executives as in the study released by Orc Worldwide based in New York on issues with HRM [3].

Found only 25 percent of managers in a systematic talent identification and most of the errors that occur when measuring talent is like measuring the wrong things, focus on the whole but not in accordance with certain talent matrix, the focus of analysis on summary data when there is hidden information that is not known and does not use data to make better decisions [4]. Talent management can be defined as a systematic and dynamic process to identify, develop and retain talent. Talent management processes are dependent on how the organization practices [5].

Methods of forecasting talent for organization employees are diverse and mostly still managed informally and through surveys of 250 respondents from the executive officers who are directly involved in the talent management of employees, the largest number of respondents using the involvement of senior leaders in talent program and a lot of using the human resource technology, the rest of them using the award based on performance, through surveys and training of personnel organization to identify and develop senior talent [6]. The technology used like decision support systems [7] [8], data mining [9, 10] and another method [11]. Employee talent can be predicted with past information available. The knowledge gained will facilitate the management of HRM and choose workers according to performance standards to avoiding the inconsistency in decision-making appointments [12].

This study gives an example technique on data preparation step, exploration of 15 data sets of selected university human resource database. 14 classification algorithms are involved in preparing the academicians talent forecasting model. The comparison results of these 14 algorithms are shown as a side conclusion from the study.

II. PREDICT ACADEMIC TALENT USING DATA MINING

Talent management in the academics is quite lacking as compared to other organization's talent [13-16]. Although there is increasing from year to year the

number of studies in the field of HRM data mining approach [17]. The studies on HRM domain from 1990 to 2011, only seven of the 106 study was associated with the talent of employees[17]. However, from seven studies that employee talent is the result of four from the same author.

Academic talent Malaysian public universities are measured in terms of professional qualifications, awards & recognition and administration & contribution to the university and identified this measure academic talent through a number of criteria in terms of employee talent Practices determination in areas such as Project Leader Assessment, Management & Professional, Academic Workers, other universities, how each of these Practices set out the criteria according to the needs of talent their respective fields [12]. While the example of other studies, academic talent as measured from an educational background, professionalism, age, gender, occupation and level of the position [18].

This study will explain in detail how the data preparation process and the experiment carried out resulted in a prediction model of academic talent management high precision after passing through several stages of the experiment. This paper is organized in five sections. The first part is the introduction, followed by the second section predict academic talent using data mining, developing academic talent, experiment result and conclusion.

III. DEVELOPING ACADEMIC TALENT

Methodology of data set preparation in this study could be summarized as the figure 1. The data preparation involves collecting 15 set raw data from 7 different systems including personnel information systems, conference information systems, university performance system, university research system, publication system, awards system and student information system. The next steps are understanding the data, import and create data marts of selected talents attributes, pre-processing such as cleaning , discretization and finally split the clean data set in the different interval year and number of classes before the experiment with classification algorithms.

A. Data Preparation

The process starts with a collection of 15 sets of raw data from the human resource database of seven systems in one selected university. The ER diagram of the 15 data sets is shown in the figure 2 below. The Demography is the main data set which consists of all lecturer id's and basic profile information. Each data set is linked with the lecturer id's.

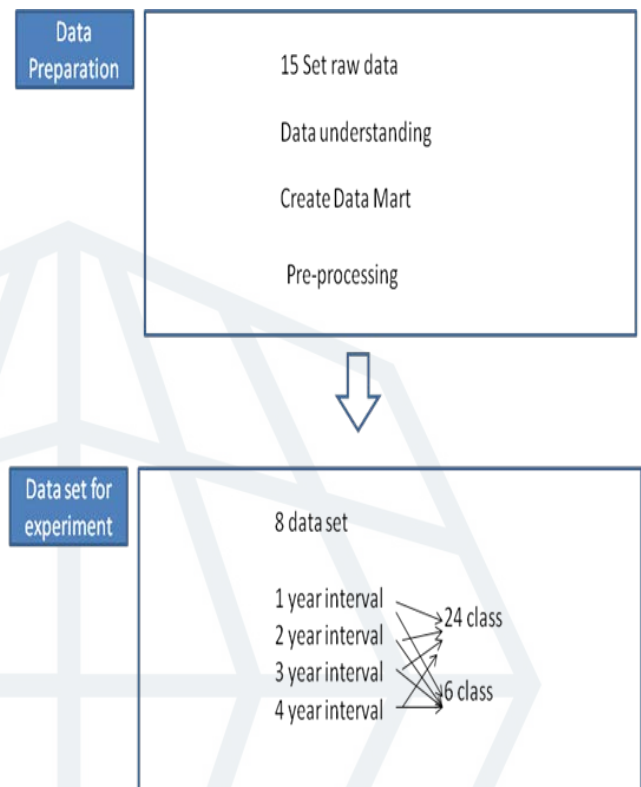


Figure 1: Data set preparation



Figure 2: ER Diagram Human Resource

B. Data understanding

To select the meaningful attribute, need to understand the pattern of records first, for example to know the distribution of received data. For each of the 15 data sets are calculated on every unique record for reason finding the meaningful attributes. For example in publication data sets, have 3 field like year of publishing, type of publications and id writers. To

create an attribute publication by year, only the year publication has a record are selected as attribute and the ways to calculate that record, using the SQL query. Here's an example query which is used to calculate each attribute value of status field from demography data set.

```
SELECT [Demography].[status],
COUNT(NZ([Demography].[status])) AS counting FROM
[Demographyi] GROUP BY [Demography].[status];
```

Analysis for data understanding, is about of 1140 attributes have been identified for talent academic attribute.

C. Create Data Mart

Data mart is a smaller scale of the data warehouse. 15 sets of raw data are transferred to the database for the purpose of creating a data mart to ease the search and integration process. This study uses Microsoft Access as the database data mart. On every attributes value needs to identify, a query is made with related data set using join query, select query and others query methods. Result from the query will be used back on programming to find and calculate automatically using that program. The figure 3 can illustrate this technique.

The program can be categorized into two types, involving only search / update data and generate the values by conducting calculations. Table 1 shows the store procedure name creating to generate the data mart by selecting the meaning full attributes from the table shows in the ER diagram.

Algorithms to update the attribute value

Results from the query

UKMPER	ID PENERBITAN	TAHUN TERBIT	KD JENIS	KIRATAHUN
K000086	TST20101360	2010	04	1
K000126	T1520092173	2006	05	1
K000126	T1520092244	2006	05	1
K000126	TST20071172	2006	05	1
K000126	TGG2008246	2008	12	1
K000126	TGG2009561	2008	05	1
K000126	TST20081664	2008	05	1
K000126	TST2009962	2008	12	1
K000126	T1520092262	2009	02	1
K000126	T1520092273	2009	02	1
K000126	T1520092275	2009	02	1
K000126	T1520092284	2009	02	1
K000126	T1520092286	2009	02	1
K000126	T1520092288	2009	02	1
K000126	T1520092507	2009	99	1
K000126	T152010121	2009	12	1
K000126	TGG2009235	2009	12	1
K000126	TGG2009652	2009	02	1
K000126	TST2010107	2009	05	1
K000126	TST2010282	2009	05	1
K000126	TST201053	2009	05	1
K000126	TST201056	2009	05	1
K000126	TST201082	2009	05	1
K000126	TST201095	2009	05	1
K000126	TST201097	2009	05	1

Figure 3: Attribute value update technique

Table 1: Talent attributes

No	Attribute involves (total of 1140 attributes)	Category	Program name
1	Idkey, year of service, position status, university status, gender, race,current position (as class)	Search / update	Demography
2	Accumulated publication (1982, 1986-2011)	Calculation	Publication
3	Number of students supervised (93/94 - 2011/2012) – did not refer the student id	Calculation	Studentbysession
4	Exact number of students supervised (93/94 - 2011/2012) – with different student id	Calculation	Studentsupervised
5	The date appoints DJJK ,VK0501 etc	Search / update	Historyposition
6	Duration from previous position to next position DJJK , VK0501 etc	Search / update	Calculationposition
7	Number of publications by year and type (publication code from 1-23 and year 1982 - 2011)	Calculation	Publication_type
8	Performance scores year (1977, 1988 – 2010)	Search / update	Performance
9	Accumulated by position on research (research positions code 1 - 9)	Calculation	Research
10	Accumulated by position on research (1 to 9) and by year (2000 – 2011)	Calculation	Research_basic
11	Number of position attending conference (1-8, A-G) by year (1996 - 2011)	Calculation	Conference_position
12	The accumulated number of attending International, Departmental, Nasional and University category	Calculation	Category_conference
13	Number of received awards by year (1990-2010) and type (Service award, Publication award and Research award)	Calculation	Award
14	Accumulated holding the administration by position (Associate members, Webmaster etc)	Calculation	Administration
15	Latest education	Search / update	Education
16	The amount of the grant received by year (1996-2011)	Search / update	Grant

Sample pseudo code for calculation publication by year, which calculates the number of publication occurs in that year:

Sub Publication ()

Declaration database

Declaration first Recordset and second Recordset

Declaration CountyYear variable

Set starting value for CountyYear as 0

Open Recordset as first Recordset (query data set)

Open Recordset as second Recordset (to store attribute value)

Open first Recordset and read next if not empty

Within the first Recordset

loop read second Recordset

Compare id key at first Recordset equal to id

key second Recordset

if equal

using case statement to check the

year of publication

if got the record, sum the

CountyYear with 1

Update value of CountyYear

into second recordset.

Looping until end of record second Recordset and first Recordset

Close first Recordset and second Recordset

Close database

End Sub

D. Pre-processing

Pre-processing is an important step in the data mining process. This study majority of the attributes is calculated based, the attribute involves filling the missing value for attributes gender, race and university status. The process is done manually by counter check others values that related to missing value for example the race attribute counter check with name familiar race and spouse information.

E. Data set creation

The overall 1140 attributes have been collected only part of the record has a value greater and equal to 1 is less than 20%. Most attributes values are 0 on the attributes of for year 2000 and below. No records were removed because if it is made, the number of attributes and records are too small and not suitable for modeling, in order to overcome this problem this study are adding the appropriate attributes related to the intervals of 2, 3 and 4 year for the attributes on category 2,3,4,7,10, 11and 13 as Table 1. The class also category to 24 classes and 6 classes as Table 2.

The 8 data sets involve being splits as below :

- Model set 1 year interval 24 classes: 3220 records and 1108 attributes.
- Model set 2 year interval 24 classes: 3220 records and 624 attributes.
- Model set 3 year interval 24 classes: 3220 records and 459 attributes.
- Model set 4 year interval 24 classes: 3220 records and 371 attributes.
- Model set 1 year interval 6 classes: 3220 records and 1103 attributes.
- Model set 2 year interval 6 classes: 3220 records and 609 attributes.
- Model set 3 year interval 6 classes: 3220 records and 454 attributes.
- Model set 4 year interval 6 classes: 3220 records and 366 attributes.

Table 2: Class attributes

DESCRIPTION	NUMBER OF CLASS		
	24	6	4
LECTURER (JKK)	A	A	A
TRAINEE DENTAL LECTURER DUG45	B	A	A
TRAINEE MEDICAL LECTURER DU45	C	A	A
DENTAL LECTURER DUG45	D	A	A
DENTAL LECTURER DUG51	E	B	B
DENTAL LECTURER DUG53	F	C	C
DENTAL LECTURER DUG54	G	C	C
MEDICAL LECTURER DU1	H	A	C
MEDICAL LECTURER DU2	I	A	A
MEDICAL LECTURER DU45	J	A	A
MEDICAL LECTURER DU51	K	B	B
MEDICAL LECTURER DU52	L	B	B
MEDICAL LECTURER DU53	M	C	C
MEDICAL LECTURER DU54	N	C	C
UNIVERSITY LECTURER DS1	O	A	C
UNIVERSITY LECTURER DS2	P	A	A
UNIVERSITY LECTURER DS45	Q	A	A
UNIVERSITY LECTURER DS751	R	B	B

UNIVERSITY LECTURER DS52	S	B	B
UNIVERSITY LECTURER DS53	T	C	C
UNIVERSITY LECTURER DS4	U	C	C
PROOFESSOR VK07	V	D	D
PROFESSOR VK06	W	E	D
PROFESSOR VK05	X	F	D

- A – Represents Lecturer
 B – Represents Senior Lecturer
 C – Represents Associate professor
 D – Represents Professor v7
 E – Represents Professor v6
 F – Represents Professor v5

F. Model Development

The forecasting model development process is based on CRISP-DM standard process which is involved choosing the technique, planning the experiment, develop model and evaluate model[19]. Figure 4 shows the development stages involved. This study was selected of 14 classification algorithms from the five main groups classification algorithm they are: J48 Decision Tree (C4.5 version 8), REPTree, Decision Stump, Random forest, Random tree, BayesNet, Naive Bayes, MultilayerPerceptron, RBFNetwork, K-star, IBK, IB1, Jrip, PART available in Weka one of the popular data mining software[20]. The discretization technique is Entropy-based & MDL stopping Criterion available in Weka for the 1st phase experiment. We choose many classifier in experiment setting to test many classifier which is to find that the most fit to identifying the academic talents.

Data sets that produce the best model from the first phase of modeling will go through 2nd stages of the experimental sessions using the discretization techniques available with different software Tanagra. Tanagra one of the best open source for the pre-processing availability technique provided[20]. 3 types of discretization techniques were chosen: Entropy-based & MDL stopping Criterion, Equal Width and Equal Frequency. Redo the experiment again with the best algorithms from earlier stages after 3 technique discretization applied to the data set. A comparison is made whether the model is with different discretization techniques and using different software for discretization can produce better result.

3rd phase, data sets from the best model will undergo the final phase of refining the model. The professor post of VK7, VK6 and VK5 merge together as show in Table 2. Then the number of classes will be reduced from six classes to four classes for further experiments. The data set is divided into three sets of data such as, the best data set with 4 classes, the best data set without interval year and 6 classes, the best data set without interval year and 4 classes.

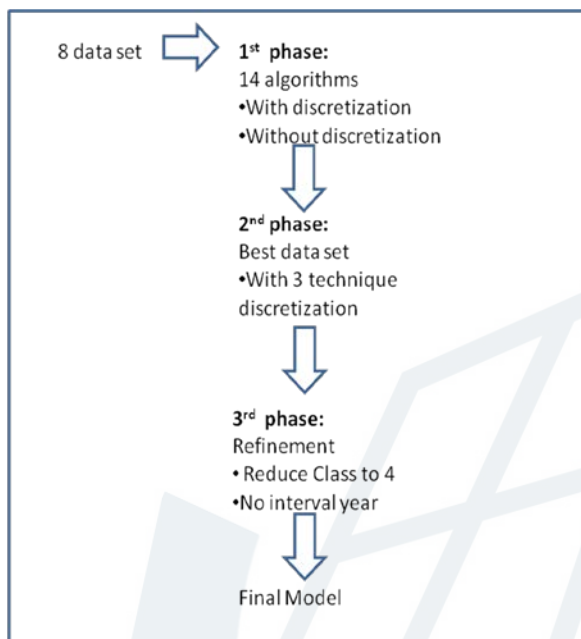


Figure 4: Model development step

IV. EXPERIMENT RESULT

Table 3 shows the overall experiment results. At the 1st phase, the best result is J48 (86.96%) with data set 4 years interval and 6 classes. Based on the results the discretization applied in the 1st phase did not show a significant difference to the results to all 14 algorithms. A conclusion also can be made , beside J48, Multilayer Perceptron, Jrip and PART also possible can produce the best prediction academicians talents as shown in Figure 5.

2nd phase experiment continues using J48 algorithms and the same data set from the best result of the 1st phase, the results did not show a significant difference with different discretization technique applied. The last phase which is 3rd phase the 4 year interval with a number of class reduce to 4 gives the best result as 87.47% and finally accepted as a talent academic model.

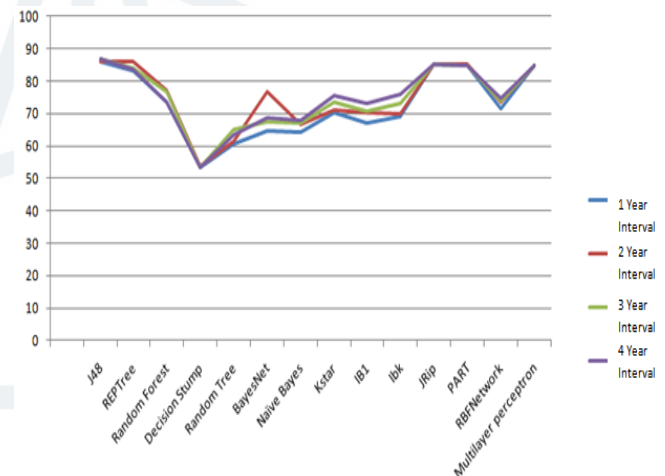


Figure 5: Pattern result from 14 classification algorithms

That shows a decision tree J48 still can give the best result compares to other classifier for talents domain same as previous research as mention in this paper. There are 117 rules successfully extracted from the best model for Lecturer, Senior Lecturer, Associate Professor and Professor talent model.

Table 3: Experiment results

Phases	Algorithm / Discretization/ Data set	Result
1st Phase	J48	86.96
	REPTree	83.33
	Random Forest	78.73
	Decision Stump	53.26
	Random Tree	65.84
	BayesNet	68.83
	Naïve Bayes	73.91
	Kstar	58.70
	IB1	67.86
	Ibk	68.32
	JRip	85.51
	PART	85.71
	Multilayer perceptron	81.10
	RBFNetwork	72.21
2nd Phase -J48 and same data set the best result (86.96)	<i>Entropy-based & MDL stopping criterion</i>	86.96
	<i>Equal width</i>	79.50
	<i>Equal frequency</i>	71.27
3rd Phase -J48 and same data but different class and interval year the best result (86.96)	4 year interval , 4 class	87.47
	No year interval, 6 class	86.34
	No year interval, 4 class	87.37

V. CONCLUSIONS

This paper has presented sample pseudo code to develop an automated academic talent data mart. The pseudo code can be stored as a store produced, to generate the latest data mart. Using the data mart, the latest academic talent model can be generated. The experiment result shows that J48 has outperformed compare to other 14 classification techniques. The result shows that applying discretization do not significantly get the better result. However, changing the number of class and arrangement in various interval years has influenced to get better results. This paper contributes on how to develop latest and accurate academic talent management using data mining and what data mining techniques and improvement to obtain better results. The accurate result is considered lower, it may because un-balance data where many zero value gets from auto generated using pseudo code.

VI. REFERENCES

- [1] H. H. A. Talib and K. R. Jamaludin, "Aplikasi Teknologi Maklumat (IT) Dalam Pengurusan Organisasi : Sorotan Kajian," *Jurnal Teknikal dan Kajian Sosial Jilid 1*, pp. 89-105, 2003.
- [2] M. Armstrong, "A Handbook of Human Resource Management Practice 10th Edition," pp. 1-957, 2006.
- [3] A. Mehta, "Human Capital Management: A Comprehensive Approach to Augment Organizational Performance," *Review of Management, Vol. 1, No. 2, April-June 2011 ISSN: 2231-0487*, vol. 1, pp. 44-57, 2011.
- [4] P. M. Powell, *et al.*, "Talent Management in the NHS Managerial Workforce," *National Institute for Health Research (NIHR)*, pp. 1-216, 2012.
- [5] B. Davies and B. J. Davies, "Talent management in academies," *International Journal of Educational Management*, vol. 24, pp. 418 - 426, 2010. doi:10.1108/09513541080000452
- [6] T. Perrin, "Talent Management: The State of the Art," *A TP Track Research Report*, pp. 1-17, 2005.
- [7] Q. Shi and M. Chen, "Design and Development of Management System for Reserve Talents of Volleyball Athletes," in *Multimedia and Information Technology (MMIT), 2010 Second International Conference on*, 2010, pp. 151-154.
- [8] L. Shi and Q. Bai, "Design a New Coherent Framework for Human Resource Personnel Evaluation Information System Based on Tasks Management," *International Conference on Business Computing and Global Informatization*, pp. 479-481, 2011. doi:10.1109/BCGIN.2011.126
- [9] C. F. Chien and L. F. Chen, "Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry," *Expert Systems with Applications*, vol. 34, pp. 280-290, 2008. doi:10.1016/j.eswa.2006.09.003
- [10] H. Jantan, *et al.*, "Classification and Prediction of Academic Talent Using Data Mining Techniques Knowledge-Based and Intelligent Information and Engineering Systems." vol. 6276, R. Setchi, *et al.*, Eds., ed: Springer Berlin / Heidelberg, 2010, pp. 491-500.
- [11] C. Mulin and H. Reen, "Arkadin develops employee talent through e-learning," *Strategic HR Review*, vol. 9, pp. 11 - 16, 2010.
- [12] H. Jantan, "Framework of Intelligent Decision Support System for Talent Management," p. 286, 2011.
- [13] C. Chen-Fu and C. Li-Fei, "Using Rough Set Theory to Recruit and Retain High-Potential Talents for Semiconductor Manufacturing," *Semiconductor Manufacturing, IEEE Transactions on*, vol. 20, pp. 528-541, 2007.
- [14] V. Mohanraj, *et al.*, "Intelligent Agent Based Talent Evaluation Engine Using a Knowledge Base," in *Advances in Recent Technologies in Communication and Computing, 2009. ARTCom '09. International Conference on*, 2009, pp. 257-259. doi:10.1109/ARTCom.2009.214
- [15] N. Goonawardene, *et al.*, "A neural network based model for project risk and talent management," *Advances in Neural Networks, LNCS 6064, Springer*, pp. 532-539, 2010.
- [16] S. Qing and C. Mengzhong, "Design and Development of Management System for Reserve Talents of Volleyball Athletes," *Second International Conference on MultiMedia and Information Technology*, pp. 151-154, 2010. doi:10.1109/MMIT.2010.59
- [17] F. Piazza and S. Strohmeier, "Domain-Driven Data Mining in Human Resource Management: A Review," in *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, 2011, pp. 458-465. doi:10.1109/ICDMW.2011.68
- [18] Y. Peng, "The decision tree classification and its application research in personnel management," in *Electronics and Optoelectronics (ICEOE), 2011 International Conference on*, 2011, pp. V1-372-V1-375.
- [19] R. Wirth and J. Hipp, "CRISP-DM: Towards a standard process model for data mining," *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, pp. 29--39, 2000.
- [20] X. Chen, *et al.*, "A survey of open source data mining systems," presented at the Proceedings of the 2007 international conference on Emerging technologies in knowledge discovery and data mining, Nanjing, China, 2007.

How to cite

Mahani Saron, Zulaiha Ali Othman, "Academic Talent Model Based on Human Resource Data Mart". *International Journal of Research in Computer Science*. 2 (5): pp. 29-35. September 2012. doi:10.7815/ijorcs.25.2012.045